Building Geoscience Semantic Web Applications Using Established Ontologies

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Matthew S. Mayernik[a], M. Benjamin Gross[b], Jon Corson-Rikert[c], Mike Daniels[a], Erica Johns[c], Huda Khan[c], Keith Maull[a], Linda R. Rowan[b], Don Stott[a]
[a]*National Center for Atmospheric Research, University Corporation for Atmospheric Research, P.O. Box 3000, Boulder, CO, USA, 80307-3000*
[b]*UNAVCO, 6350 Nautilus Drive, Boulder, CO, USA 80301-5394*
[c]*Cornell University Libraries, Cornell University, Ithaca, NY, USA, 14853*

Abstract. Interplays between local ontology development and the establishment of wider ontology connections are fundamental to the Semantic Web. This paper discusses the goals and work of the EarthCollab project, focusing on ontology selection, consolidation, and reuse driven by geoscience use cases. The EarthCollab project is a collaboration between UCAR, Cornell University, and UNAVCO to leverage semantic technologies to manage and link geoscientific information and resources. EarthCollab is using the VIVO Semantic Web software suite to support the discovery of information, data, and potential collaborators within the geodesy and polar science communities. This paper presents the EarthCollab ontology design approach, which is heavily emphasizing ontology reuse, and discusses how the different needs of each use case have informed our ontology selection and design. The EarthCollab project is bringing together the VIVO-Integrated Semantic Framework (VIVO-ISF) ontology, the Global Change Information System (GCIS) ontology, and the Data Catalog (DCAT) ontology, among others, to support diverse use cases related to the discovery of geoscience information and resources. Understanding the challenges and solutions for ontology reuse are critical to informing key decision points for new semantic web applications in deciding when to reuse existing ontologies and when to develop original ontologies.

Keywords: ontologies, ontology reuse, geoscience, VIVO

1. Introduction

The interplay between local ontology development and the reuse of existing ontologies is fundamental to the Semantic Web. The openness and flexibility of the Semantic Web, allowing "Anyone to say Anything about Any topic" (the AAA slogan), is a designed feature of the underlying data representation layers (e.g. RDF and RDFS). The Semantic Web is rooted in the principle of universality, allowing anything (digital or physical) to be considered a resource and assigned a URI. The principle of the "open world" is another key component of the Semantic Web, stipulating that new statements about Semantic Web resources are always possible, and that there should be no assumption that a URI uniquely refers to an individual resource [14]. Building a web of data on the internet, however, requires some consistency in how data are represented in order to build a meaningful network. As stated in an early description of the Semantic Web by Berners-Lee, Hendler, and Lassila [2]: "The Semantic Web, in naming every concept simply by a URI, lets anyone express new concepts that they invent with minimal effort. Its unifying logical language will enable these concepts to be progressively linked into a universal Web." The two sentences in this quote illustrate the tensions between local ontology development and ontology reuse. Anyone can invent and express new concepts, but the "universal Web" can only emerge through tying concepts together.

This paper explores the effort of efficiently developing a local ontology while employing current community ontologies within the context of the EarthCollab project and geoscience Semantic Web applications more broadly. The EarthCollab project is a collaboration between the University Corporation for Atmospheric Research (UCAR), UNAVCO, and Cornell University to use Semantic Web and linked data technologies to facilitate the coordination and organization of complex scientific projects, their communities, their tools, and their products. EarthCollab is funded by the National Science Foundation's EarthCube initiative, which supports the development of cyberinfrastructure for the geosciences. EarthCollab's aim is to enable researchers to more easily find people, organizations, and research resources that are relevant to their work. The EarthCollab partners are using the VIVO semantic software suite to enable more coherent discovery of distributed information and data for large multidisciplinary scientific projects. From an ontology point of view, the EarthCollab preference and goal has been to reuse existing ontologies as much as possible to facilitate easier integration and data sharing across the geoscience community.

The main questions this paper addresses relate to our efforts to reuse established ontologies from inside and outside the geosciences domain in the development of our EarthCollab ontology and web application. The questions include:

- What are the key decision points for new Semantic Web applications in deciding when to reuse existing ontologies and when to develop original ontologies?
- How can new Semantic Web projects most efficiently and effectively identify and select ontologies to reuse?

Through this discussion, this paper contributes to the understanding of how to develop and manage applications that are based on multiple ontologies.

2. Background

Constructing original ontologies and reusing existing ontologies present challenges to Semantic Web application developers. The focus of this paper, reuse of existing ontologies, has been discussed from a number of angles. Introductory texts on ontology development often list the ability to reuse existing ontologies as one of the key attractions of the Semantic Web. As one such introductory text states: "Why build something if it is already available on the Web? One of the easiest ways to begin a [Semantic Web] modeling project is to find models on the Web that suit your needs" [1]. Adopting existing ontologies for a new application, however, requires substantial knowledge of Semantic Web principles and underlying technologies. Naïve approaches, such as simply selecting ontologies that match keywords of interest, can lead to "Frankenstein ontologies" that run afoul of good ontological engineering practices [3].

Semantic Web methodologies recommend that ontology design and selection should follow from application-specific concept designs [11,13]. At a conceptual level, reusing ontologies thus requires evaluating representational models and task-specific assumptions and decisions [36]. Existing ontologies may not fully support new applications, or may have a greater or lesser level of complexity and/or granularity than required for the new application. Small differences in the tasks being supported by different applications can lead to substantially different requirements for ontological models.

Previous literature on ontology development includes several discussions and recommendations on how to properly import one ontology into another, describing the technical aspects of reusing portions of existing ontologies. The "Minimum Information to Reference an External Ontology Term" (MIREOT) recommendations discuss a method for importing information from existing ontologies into new ones [5]. The MIREOT recommendations are meant to address the selective reuse of a subset of terms from a larger ontology. The MIREOT approach has been most influential in biomedical-related ontology development (see for example [19,35]). Other approaches to support the use of subsets of existing ontologies have also been proposed, including semantic import [29] and ontology modularization [8].

Even with these approaches, overcoming task and domain-specific ontology dependencies is a considerable challenge [16]. Negatives of improper ontology reuse come in a number of forms. Hogan, Harth, and Polleres [17] describe the potential for ontology reuse to lead to "ontology hijacking," where the secondary ontology can introduce ontological commitments that change the semantics of specific classes or properties in the original ontology. For example, declaring new superclasses of classes from commonly used ontologies, such as foaf:Person, can introduce problematic inferences for other applications that use those same classes. Similarly, inconsistent use of the owl:sameAs property can result in inferencing of

equivalences between two entities that may not be appropriate in all contexts [15].

The full semantic implications of ontology reuse or recombination may only be visible through detailed comparison of conceptual models, encompassing a close reading of definitions and the full range of axioms, and not just a comparison of the class and property hierarchies. Conceptual models that use differing underlying higher-level ontologies may not be compatible, even if the same concepts are being modeled. In a pair of studies, Cox [6,7], for example, analyzed two different approaches to modeling "observations" within scientific contexts. The use of different higher-level ontologies led to "observations" being modeled as "events" in one approach versus as "objects" in another approach. Both of these uses of the term "observations" might be appropriate and well-defined for their specific contexts, but combining statements about "observations" from each context might lead to inconsistencies and problems in interpretation.

3. Geoscience Semantic Web Applications

The use of the Semantic Web in geoscience information and data applications ranges widely, and includes development of ontologies at many points along the ontology spectrum [27], from controlled thesauri to formal ontology specifications with properties, logical constraints, and underlying axioms. The Semantic Web for Earth and Environmental Terminology (SWEET) ontologies were born from an initiative to define a knowledge space for Earth system science and related concepts [31]. They model key concepts within Earth system science, such as space, time, Earth realms, physical quantities, as well as specific scientific concepts and phenomena. Other projects focused on more specific applications. The development of the Virtual Solar-Terrestrial Observatory (VSTO), for example, included the creation of an extensive ontology that models solar-related phenomena and scientific observational tools and procedures [12]. In addition, representational models from the Open Geospatial Consortium (OGC) have been converted into Semantic Web data structures to enable wider use by Semantic Web-based applications [6,30].

The SWEET and VSTO ontologies were original ontologies that modeled their respective concepts and relationships in considerable detail. SWEET, for example, contains 6000 concepts in 200 separate ontologies. Subsequent geoscience-based Semantic Web projects swung toward ontology reuse, though original ontology development is still a feature of many projects. Recent efforts that mix original ontology development with ontology reuse include:

- Global Change Information System (GCIS) - An ontology-based collection of information related to climate and environmental change [26]. Reused ontologies include the Dublin Core Type, Organization, Friend of a Friend (FOAF), and the Bibliographic Ontology (BIBO). In addition, the GCIS ontology bases most of its structure on the W3C's Provenance Ontology (PROV-O).
- The LASP Extended Metadata Repository (LEMR) - A semantically enabled repository of information about datasets managed and made accessible by the Laboratory for Atmospheric and Space Physics [37]. LEMR's ontology incorporates components from the FOAF, BIBO, and Data Catalog (DCAT) ontologies, among others.
- Sea Ice Semantics - An initiative to create interoperable ontologies that incorporate knowledge of sea ice from multiple perspectives, including knowledge derived from academic scientific studies, from practices related to forecasts shipping lanes, and from residents of communities that live in the Arctic [9]. This effort incorporates concepts and relationships from the SWEET ontologies, and from the taxonomies found in the National Aeronautics and Space Administration's (NASA) Global Change Master Directory (GCMD).
- Deep Carbon Observatory (DCO) - The DCO is developing a set of portals to enable the discovery and use of information and data related to the cycle of carbon through the entire Earth system [24]. The DCO ontology incorporates aspects of the VIVO-Integrated Semantic Framework (VIVO-ISF), as well as the DCAT and other ontologies.

Another set of projects, called OceanLink and GeoLink, have taken a different approach to modeling ocean and other geoscience information, emphasizing original ontology development using an Ontology Design Patterns approach [22,23]. These projects are developing a set of granular ontology design patterns for a range of phenomena and processes, such as for modeling oceanographic cruises, in order to create a common set of representations of resources held across numerous geoscience information and data providers.

4. EarthCollab Technology

EarthCollab is using the VIVO open-source software suite to represent and describe scientific networks (http://vivoweb.org). Since 2003, VIVO has pioneered a semantic approach to modeling research and scholarly activities focused on connecting many different types of entities – people, organizations, events, courses, grants, and publications – through named relationships. The primary use of VIVO is to provide an infrastructure for research networking, leveraging personal profiles, publication records, grant information, and subject expertise to enable the discovery of research and scholarship across disciplines. A sample VIVO profile is shown in Figure 1. This is a customizable visual display of information about a faculty member, with publication, research, and other information provided in one location.

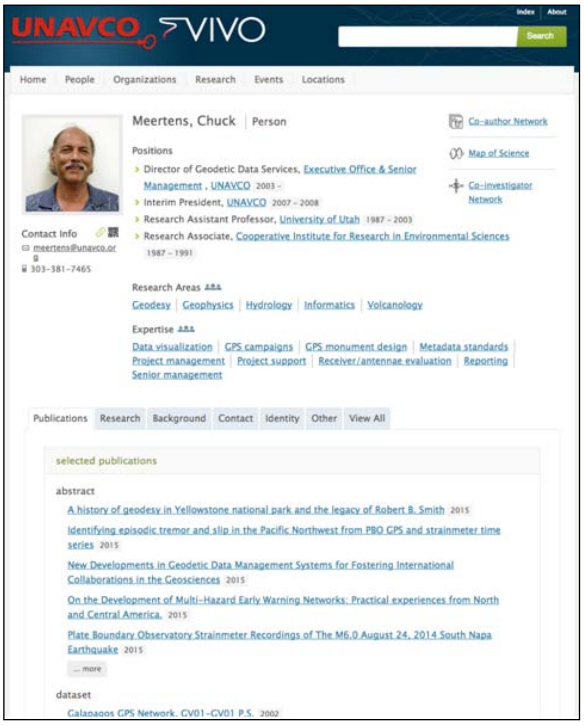

Fig. 1. Sample person profile in the UNAVCO VIVO application.

The EarthCollab project is extending the use of the VIVO software to scientific settings, representing datasets, scientific instruments, and research projects. Figure 2 shows an example of a dataset profile from the UNAVCO VIVO instance. Datasets have associated publications, organizations, creators, managers, instruments, and subsequent datasets. VIVO has emphasized from its inception the linked data principle, such that even though people are powerful vectors of connection, other resources provide substantial merit for discovery and connectivity, and should be treated as objects in their own right. Making network connections explicit, bidirectional, and visible adds context, supports multiple points of access, and provides paths of navigation through data. Within the VIVO data model, almost everything can be represented as a first-order object – not just people, organizations, publications and grants, but instruments, projects and their components, work groups, datasets, methodologies developed, presentations, and any other items of interest declared using appropriate ontologies. EarthCollab is one of a number of projects that are using VIVO to represent scientific information. Two of the geoscience projects noted above, the Deep Carbon Observatory and Laboratory for Atmospheric and Space Physics (LASP) initiatives, are using VIVO within their respective technology stacks.

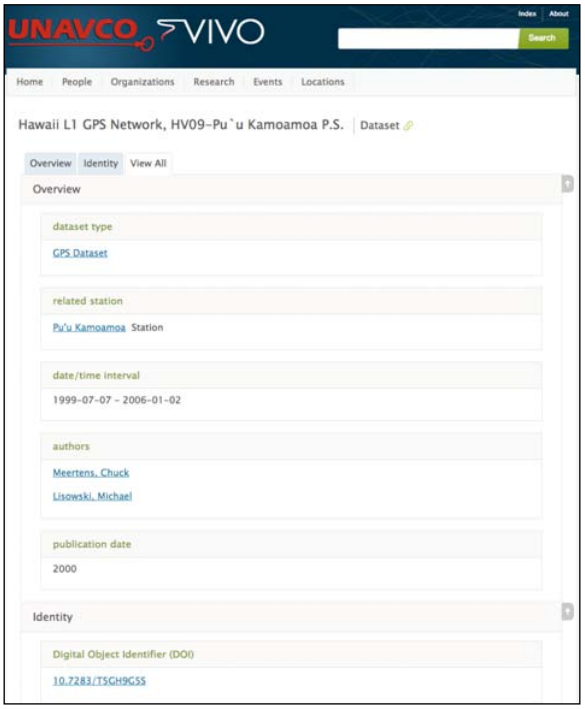

Fig. 2. Sample dataset profile in the UNAVCO VIVO application.

The VIVO software suite leverages the VIVO-ISF (Integrated Semantic Framework) ontology. The VIVO-ISF ontology, discussed more below, defines types (classes), and the relationships between them (properties), for researchers, organizations, and a range of research activities, such as publications [28].

Existing widely used ontologies are reused in the core VIVO ontology, such as Friend of a Friend (FOAF) to define people, ensuring that VIVO can easily exchange and integrate data with external data sources. In addition, the inherent extensibility of the VIVO application enables the integration of any RDF/OWL ontology [21].

5. EarthCollab Use Cases

The two main stakeholder communities for the EarthCollab project are: (1) the Bering Sea Project, an interdisciplinary field program whose data archive is hosted by NCAR's Earth Observing Laboratory (EOL), and (2) UNAVCO, a geodetic facility and consortium that supports diverse research projects informed by geodesy. The interests of these communities have guided our process of deciding which specific ontologies to investigate and reuse. To develop a better understanding of our stakeholder community, and their uses for linked data, EarthCollab conducted a set of use case exercises.

Use case development exercises were conducted within the NCAR EOL and UNAVCO data center teams to compile high-level statements of specific tasks that scientists should be able to achieve through the use of EarthCollab systems. The use case descriptions followed the template provided by Fox and McGuinness [11]. This use case activity identified key tasks that EarthCollab systems should support, including: "finding all publications that used the GPS data held by UNAVCO", and "identifying the people responsible for the collection of a specific Bering Sea data set held by NCAR EOL." These use cases led to the development of concept and information models by identifying key concepts and terminologies that our ontology will need to model. The concept models developed based on the use cases depict the key entities of interest for our use cases, e.g., publications, data, people, organizations, instruments, projects, etc. The concept models also depict the kinds of relationships that can exist between these entities.

In addition to the data center-focused use cases, EarthCollab conducted a user engagement workshop in December of 2014 to develop science-focused use cases. Nine scientists (three related to the Bering Sea Project and six related to UNAVCO) participated in the workshop, and were led through use case development discussions. The resulting use cases had a couple of different focuses. One focus was an interest in identifying geospatial regions where certain data parameters show specific features, such as "Find the areas in a defined region where benthic biomass is high enough for walrus feeding". Another use case focused on finding information (e.g. data, publications, or other products) related to a geolocated event, such as a seismic event or a scientific field project. Based on these use cases, important areas for ontology development included the modeling of geospatial and temporal features of research projects and their resulting data sets. Other ontology priorities that came out of the workshop were the modeling of project-specific information, e.g. ship tracks for oceanographic cruises or stations and instruments that are part of a given UNAVCO GPS network.

With these use cases in hand, NCAR EOL and UNAVCO have each set up an installation of the VIVO system, and are actively ingesting information about the people, publications, and data sets within their respective communities. Ingested information comes from existing metadata databases at NCAR EOL and UNAVCO, and from newly developed sources.

6. Key Ontologies being (Re)used in EarthCollab

This section describes the three ontologies identified as being most appropriate for EarthCollab applications: (1) VIVO-ISF, (2) GCIS, and (3) DCAT ontologies. Each ontology is briefly described, highlighting their key characteristics and relevance for EarthCollab. The process used to identify and combine these ontologies is subsequently discussed in the following section.

(1) VIVO-ISF (Integrated Semantic Framework) ontology - https://github.com/vivo-isf

The VIVO-ISF ontology is primarily an ontology representing scholarly work. It focuses on modeling people, academic organizations, and their scholarly products and activities. The VIVO-ISF ontology grew out of the merging and extension of two existing ontologies, eagle-i and the original VIVO ontology, which were developed in the context of separate National Institutes of Health (NIH) grants. The goal of the merge was to bring more coherence to NIH funded ontology development initiatives. From eagle-i, the VIVO-ISF incorporates the Basic Formal Ontology (BFO) as an upper ontology. The original VIVO ontology had no upper ontology. Also, because the eagle-i ontology was developed to describe biomedical research resources with significant detail,

the VIVO-ISF has inherited terms specific to the biomedical domain, such as terms related to the phases of clinical trials [35]. The design goals in creating the merged VIVO-ISF ontology were to identify overlapping and duplicate entities between the two ontologies and to avoid severe disruptions in either VIVO or eagle-i application compatibility. The ontology supplied with the VIVO application is a subset of the merged VIVO-ISF ontology, with most anatomy and biomedical research concepts removed.

The VIVO-ISF ontology contains many original classes and properties, but also incorporates aspects of numerous other ontologies including BFO, BIBO, CL, Event, FOAF, GO, Geopolitical, IAO, OBI, OCRe, vCard and others (https://wiki.duraspace.org/x/3AQGAg). It includes object domains and ranges that have implications for the operation of the VIVO application.

As the backbone of the VIVO application, the EarthCollab project uses the VIVO-ISF ontology as the core construct for modeling people, organizations, grants, and publications.

(2) Global Change Information System (GCIS) ontology - http://data.globalchange.gov/gcis.owl

The GCIS ontology was developed by the Tetherless World Constellation (TWC) at Rensselaer Polytechnic Institute (RPI), as part of the the U.S. Global Change Research Program. The GCIS ontology underlies a web-based information system, also called the Global Change Information System (GCIS), that supports exchange of global change information between federal agencies. The GCIS system was developed to establish interfaces to interoperable repositories of climate and global change information. The initial focus of the GCIS ontology was to model entities and relationships related to the Third National Climate Assessment report (http://nca2014.globalchange.gov/). The GCIS and the GCIS ontology are intended to be expanded to support additional use cases in the future. Because tracking inputs and outputs of the National Climate Assessment are key drivers of the GCIS, the GCIS ontology uses the W3C Provenance ontology (PROV-O), an ontology for capturing provenance information, as the base ontology [25,26].

Though the terms defined in the GCIS ontology are generalized, the similarity of the project's domain with those in EarthCollab's use cases make it an obvious choice for integration in our project. Specifically, EarthCollab is using the GCIS ontology to model scientific instruments, platforms, datasets and the relationships between them.

(3) Data Catalog (DCAT) ontology - http://www.w3.org/TR/vocab-dcat/

The DCAT ontology provides a compact structure to model datasets, data catalogs, and associated information. The DCAT originated at DERI (Digital Enterprise Research Institute at NUI Galway, Ireland), and then was standardized by W3C and its Government Linked Data (GLD) Working Group. The goals of DCAT are to represent data and data exchange. As such, it includes core classes such as Catalog, Dataset, and Distribution. The DCAT ontology uses existing ontologies, such as the Dublin Core Terms ontology, to provide many class attributes.

Endorsed by the W3C, the DCAT ontology has been used and expanded a number of times [34,37] including in a recommendation by a European Commission program to standardize data portals across the European Union [10]. The EarthCollab project uses the DCAT ontology to describe datasets in greater detail than is possible with the VIVO-ISF and GCIS ontologies.

7. EarthCollab Ontology Development

This section describes the process used in the EarthCollab project to identify appropriate ontologies, and to develop a set of effective ontological structures.

7.1. EarthCollab concept & ontology modeling

EarthCollab's process to identify the appropriate ontologies included canvassing established ontologies and developing use cases as described above. These use cases helped to identify one set of key concepts and terms of interest for our ontology selection and development. Another set of key concepts, terms, and relationships were developed through looking at the metadata schemes already in place within our two data center partners, UNAVCO and NCAR EOL. These existing metadata stores contain significant amounts of structured and unstructured information with community-vetted terminologies and established relationships between entities. They imposed important constraints on the EarthCollab ontology approach, limiting the statements that can be made about datasets and their relationships to other key entities, such as investigators and observational platforms.

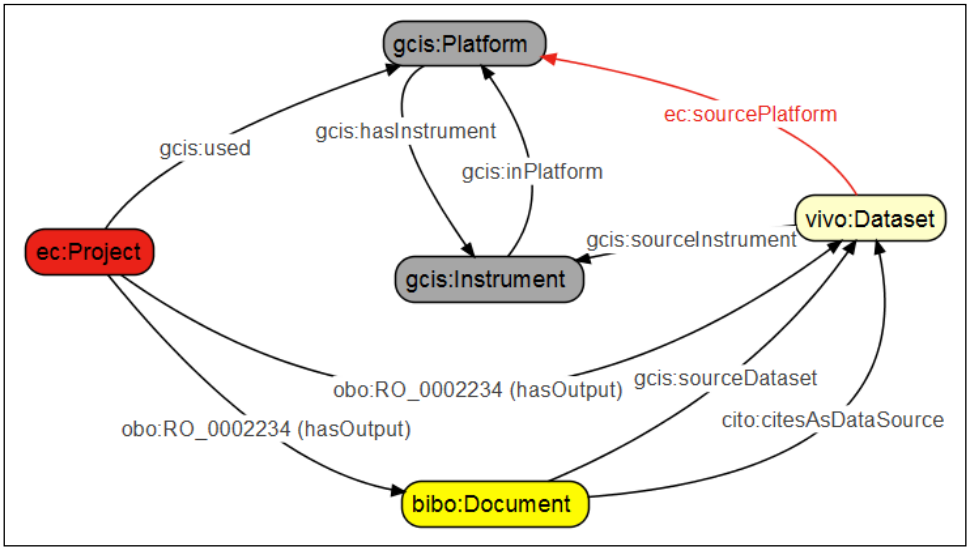

Fig. 3 - Core EarthCollab ontology concepts. Prefixes and colors indicate different underlying ontologies. Concepts and relationships shown in red are specific to the EarthCollab (EC) project.

A key part of this activity was to identify overlaps, consistencies, and differences between the UNAVCO and NCAR EOL conceptual structures. During these discussions, the EarthCollab partners often found that with minor adjustments they could conform to a common semantic dialect.

In parallel, the EarthCollab team investigated existing ontologies that mapped to the concept and relationship structures emerging from the use cases and metadata analysis. Discovering and evaluating ontologies is an informal process at present. Domain-specific ontology portals, such as http://ontobee.org or http://bioportal.bioontology.org/, are of limited utility for projects outside of those domains. General purpose ontology portals such as the Linked Open Vocabularies portal (http://lov.okfn.org/) are useful, but personal communications and networking at workshop and conference venues have proven to be more fruitful in terms of identifying ontologies of direct relevance and applicability to our project. EarthCollab project members have gathered information about potentially relevant ontologies by participating in discussions within the geoscience community, such as the Earth Science Information Partners (ESIP) Semantic Web Cluster (http://wiki.esipfed.org/index.php/Semantic_Web), the NSF EarthCube program, and the American Geophysical Union's Earth and Space Science Informatics (ESSI) focus group.

Figure 3 shows an early EarthCollab concept and ontology diagram. This diagram depicts a set of core concepts and relationships that encompass key aspects of the use cases and map to existing metadata schemes used by UNAVCO and NCAR EOL. Many details are omitted in the creation of this diagram (such as class attributes), but the diagram provided a useful starting point for clarifying ontology needs.

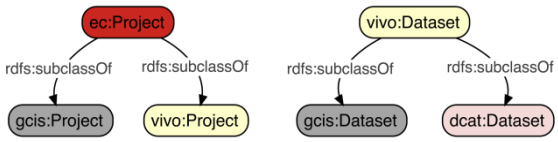

Fig. 4. Sub-class relationships between EarthCollab ontology classes

As shown in Figure 4, the next step was to specify relationships between similar classes that are present in multiple ontologies. This involved investigating whether any of the existing ontologies, such as DCAT or GCIS, contained any restrictions that might limit or preclude reuse in the EarthCollab context. In examining domain and range restrictions on properties, as well as class hierarchies, for potential conflicts, necessary EarthCollab-specific classes and properties were identified. In particular, the VIVO-ISF and GCIS ontologies contained Project classes,

and VIVO-ISF, GCIS, and DCAT ontologies all contained Dataset classes. The structures in Figure 4 were chosen to avoid "hijacking" the VIVO-ISF and GCIS Project classes, as well as to take advantage of properties associated with both. In the Dataset case, the GCIS and DCAT ontological structures have much more comprehensive modeling related to datasets than the VIVO-ISF ontology. Based on our discussions within the VIVO developer community, the Dataset class in the VIVO-ISF ontology (version 1.6, current as of this writing) has not been used extensively in VIVO applications. As such, there is little concern about causing unintended ontological commitments by defining it as a subclass of the corresponding GCIS and DCAT Dataset classes. Additional EarthCollab-specific ontology components, such as the ec:sourcePlatform property shown in Figure 3, filled gaps that none of the other ontologies covered.

Another case of overlapping terms that was less important for EarthCollab was the conflict between the dcat:Catalog and vivo:Catalog terms. In DCAT, a catalog is a "curated collection of metadata about datasets". In the VIVO-ISF ontology, the Catalog class has no textual definition, but is categorized as a subclass of vivo:Document, suggesting that it is intended to apply to published items. This conflict is not critical to address in the EarthCollab context because our use cases do not require extensive modeling or use of either "catalog" concept. In addition, because the two classes have no other ontological clashes, it was unnecessary to try to consolidate the two ontology structures formally.

Examining domain and range restrictions on object properties can reveal aspects of ontologies that are domain-specific, making it difficult to reuse them. For EarthCollab purposes, for example, it is important to model the relationships between an instrument, a technique, and a person. The GCIS ontology contains instrument and software classes and many associated properties, but does not define a relationship between those classes and people. VIVO-ISF includes object properties designed to make connections between instruments and software, but includes domain and range restrictions that are not ideal for our EarthCollab use cases. For instance, the VIVO-ISF Used by object property has a domain of software, database, and protocol and a range of Organization. The included example, "A laboratory uses Microsoft Word" attests to the eagle-i ontology's biomedical-centric design, which has been inherited by VIVO-ISF. Rather than "hijacking" VIVO-ISF by tweaking the domain and range assertions in VIVO-ISF, it makes more sense to create an EarthCollab-specific object property for modeling connections between people and the instruments, methods, or software that they have expertise to use.

As noted above, existing ontologies often include more detail than required by secondary applications. In the EarthCollab case, source ontologies were selectively trimmed to create the combined project ontology (the MIREOT approach). The GCIS ontology was edited extensively, removing a number of classes and properties that were not relevant for our use or would potentially clash with other ontology components. For example, like the VIVO-ISF ontology, the GCIS ontology models people, organizations, and their relationship to other entities. The GCIS structure, however, differs from the VIVO-ISF ontology's approach for modeling these entities. For example, the VIVO-ISF ontology models the concept of "authorship" as a class, whereas the GCIS ontology models authorship via an isAuthorOf property. VIVO-ISF's use of "context classes" [32,33], shown in Figure 5, allows for additional properties to be assigned to an entity of the Authorship class, a "context node." For example, these context classes allow the designation of author order on a particular document. In order to apply the Structure of the paper

Fig. 5. The VIVO-ISF ontology structure for modeling authorship.

The same "context node" approach is used in the VIVO-ISF ontology in modeling person and organizational roles as classes, such as MemberRole and FundingRole. As with the authorship example, these roles are modeled in the GCIS ontology as direct properties, namely hasMember and hasFundingAgency. Recognizing the advantages of "context nodes," and because the VIVO software is customized to support the VIVO-ISF ontology, the components of GCIS that focus on people and organizations were removed.

Reusing ontologies often also requires addition of entities and relationships that are needed for a new application. For example, UNAVCO and NCAR EOL are consortia of universities and related geoscience institutions. The person and organization aspects of the ontology need to represent classes of members that are external to either institution. Figure

6 shows an EarthCollab custom class, AssociateMemberRole associate member class, and a custom property, hasLiaison, which are important to the UNAVCO application in connecting the UNAVCO consortium to its member institutions and associated representatives.

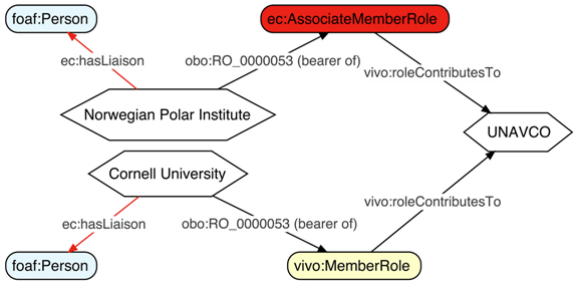

Fig. 6. Illustration of VIVO's modeling of membership, and EarthCollab's local extension to include the concept of associate members of a distributed institution.

7.2. EarthCollab ontology work in progress

Ontology approaches for other aspects of EarthCollab's use cases are still being investigated. The scientific use cases noted above have a strong geospatial component. Assessing the various options for modeling geospatial information, however, is a lengthy task, because so many approaches exist [20]. Another scientific concept that is salient in our use cases is the importance of events, such as earthquakes or hurricanes, for organizing scientific information, data, and projects. Figure 7 illustrates a model for how to fit event-related information into the EarthCollab ontology structure, with question marks showing open questions. In practice, the notion of an "event" has proven difficult to track in existing ontologies. The VIVO-ISF ontology has an Event class, which is drawn from the Event ontology (http://purl.org/NET/c4dm/event.owl). The Event ontology is primarily intended to model human events, such as conferences and concerts, but is generic enough that geospatial events could be modeled. The VIVO-ISF ontology includes a property specific to the Event class, the label for which is "related documents". Initially, this combination of class and property appeared appropriate to our case, as it could be stated that a given scientific event, such as an earthquake, had a set of related documents housed at UNAVCO or NCAR EOL. Looking closer, however, at the URI for the VIVO-ISF "related documents" property, showed that the actual underlying property is bibo:presents, a much more specific property than the label "related documents" suggests. Using this property for our geoscience case, then, becomes inappropriate, (e.g. this would effectively be saying that "a document was presented at an earthquake", an obviously false statement).

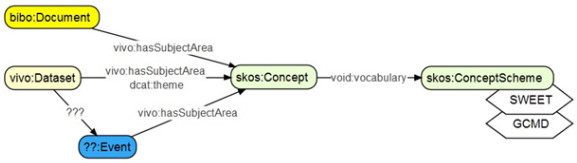

Fig. 7. Concept diagram for modeling of events and subject terms.

Another aspect of our project that requires more consideration of existing ontologies is the possible need to describe a network. UNAVCO maintains a network of field stations, with each station containing a suite of instruments. Researchers and practitioners might use data from the entire network, a part of the network, a few stations, or only from specific instruments. The "station" concept does not cleanly map to existing ontologies that EarthCollab has investigated to date. The addition of "station" as an EarthCollab defined concept is being considered, as shown in Figure 8. Another uncertainty is about whether to consider the addition of a "network" concept into the ontology. One initial idea was to use the Project class to represent a network, but in subsequent modeling, it was decided that the terms are not equivalent.

8. Discussion

It is rare that a pre-existing ontology meets all of a new application's needs. In the EarthCollab case, clearly defined use cases helped to determine the extent to which existing ontologies should be used. Inventing a new ontology is required when an application has specific needs or concepts that existing ontologies do not address. In EarthCollab, a key goal has been to use the Semantic Web to share information about datasets, projects, people, etc., within the project and externally. The project's ontology approach has thus been oriented toward enabling interoperability of data with as few barriers as possible.

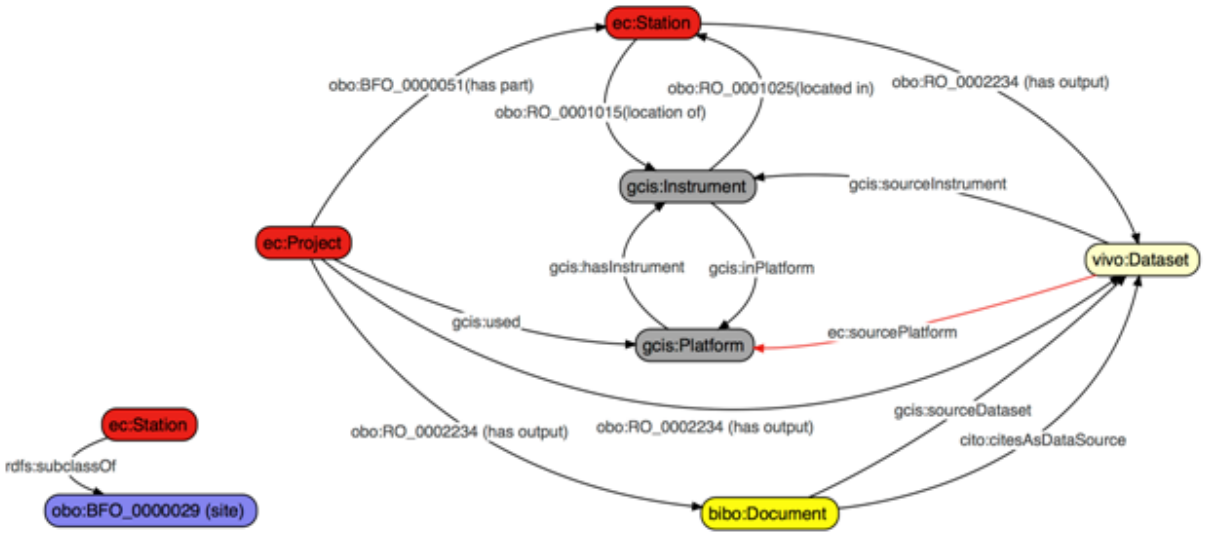

Fig. 8 - Expanded EarthCollab ontology diagram depicting the addition of the Station class and its relationships with other classes.

Reusing ontologies facilitates interoperability and data exchange. The application software used in the EarthCollab project will likely be relatively short lived. The semantic data, however, will be able to be used in additional existing and future software applications.

Ontology design might be informed by the application, or the reverse, the application might be informed by the ontology selection process. Ontology design can be distorted if it is too driven by the application, such as the need to display certain kinds of information in certain ways. The EarthCollab ontology design approach has been a combination of these paths: driven by the use cases, the metadata that UNAVCO and NCAR EOL brought to the project, and the Semantic Web application being used, VIVO. For example, the VIVO software is architected around the use of the VIVO-ISF approach that models authorship as an context class instead of as a property in order to express author order. Data input and display functions in the VIVO application are structured to work with data that use those context classes. Thus, the desire to avoid having spurious classes being displayed in the VIVO application drove the decision to cut back a portion of the GCIS ontology related to people, organizations, and their roles. Similarly, existing ontologies may lack domain and range declarations on some or all object properties. In the VIVO application, however, it is often helpful to have domain and ranges expressed through value constraint property restrictions (http://www.w3.org/TR/owl-ref/#Restriction), because they are used to structure web forms and information displays.

Maintenance and versioning are key considerations for ontology reuse [18]. Some ontologies have more robust maintenance support than others. Ontology source files, namespaces, and support pages might disappear or be moved, just as any other web-based resource. Ontology updates can be a source of improvement or problems for ontology reusers. For example, the VIVO-ISF Authorship class, discussed above, was introduced as part of an ontology update. The first VIVO ontology modeled many person and organizational roles via direct relationships [4,28,35]. This ontology version change had significant implications for the VIVO applications already deployed and using the prior VIVO ontology. VIVO application users who transitioned to the new ontology faced a sizable conversion process, and a small minority of VIVO applications users decided not to make the switch. This example illustrates how ontology versions should be available in perpetuity to allow for applications to continue to use prior versions if desired. The reality of ontology maintenance and versioning is one of the most significant challenges to ontology reuse. It may push new Semantic Web applications to limit the scope of ontology reuse, because potential versioning problems are inherently tied to the number of ontologies in use.

9. Conclusion

The EarthCollab project emphasized the reuse of existing ontologies to support the sharing of information about scientific data, projects, and researchers. Using the VIVO software suite with its built-in VIVO-ISF ontology as our base, EarthCollab has investigated and selected an additional set of existing ontologies to more fully cover the concepts and relationships relevant to geoscientific research. The process used to integrate these additional ontologies into the VIVO application has involved multiple steps, and has been done as much as possible following recommended practice for ontology import and reuse.

The next steps for our project are, in the immediate term, to conduct systematic user tests that encompass the individual VIVO instances being operated by NCAR EOL and UNAVCO, as well as mechanisms, currently under development, for exchanging and sharing information across multiple VIVO instances. The user tests will inform our continued development of our ontology structures, by helping to illustrate the type of conceptual linkages that users hope or expect the EarthCollab project to facilitate. These user tests will help the EarthCollab project to engender engagement and support by our target user communities in the geosciences, and will help to get a better understanding of the challenges in rolling out Semantic Web-based applications to researchers who are unfamiliar with Semantic Web or linked data technology. Demonstrations that illustrate the true utility of the technology to organizational leaders and decision-makers will increase the value and understanding of these systems.

Finally, as the EarthCollab project matures, the project members will continue to investigate how the ontological approach described in this paper can facilitate sharing information with additional external partners. For example, initial discussions between EarthCollab project members and members of the GeoLink project [23] have identified promising potential overlaps between the two projects, at an ontological level and in the actual data being modeled. Investigating the interoperability challenges involved in connecting these two approaches, EarthCollab focused on ontology reuse and GeoLink focused on the development of ontology design patterns, would hopefully provide unique and valuable information to future ontology connection efforts in the geosciences and broader communities.

10. Acknowledgements

EarthCollab is funded by the US National Science Foundation, grants # 1440293, 1440213, 1440181, PIs Matthew Mayernik, Mike Daniels, Linda Rowan, and Dean Krafft. In addition to the listed authors of this paper, EarthCollab project members include Fran Boler and Chuck Meertens from UNAVCO, Dean Krafft from Cornell, and John Allison, Scot Loehrer, Mary Marlino, Steve Williams, and Mike Wright from UCAR.

References

[1] D. Allemang and J. Hendler, Semantic Web for the Working Ontologist, 2nd Edition. New York: Morgan Kaufmann, 2011, pp. 307.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web. Scientific American, May 2001, pp. 34-43.

[3] O. Corcho, M. Poveda-Villalón, and A. Gómez-Pérez, Ontology engineering in the era of linked data, Bulletin of the American Society for Information Science and Technology, 41(4): 13-17, 2015. http://doi.org/10.1002/bult.2015.1720410407

[4] J. Corson-Rikert, S. Mitchell, B. Lowe, N. Rejack, Y. Ding, , and C. Guo, The VIVO ontology. In K. Borner, M. Conlon, J. Corson-Rikert, Y. Ding (Eds.), VIVO: A Semantic Approach to Scholarly Networking and Discovery, Morgan & Claypool, pp. 15-33, 2012.

[5] M. Courtot, M. Courtot, F. Gibson, A. Lister, J. Malone, D. Schober, … and A. Ruttenberg, MIREOT: the Minimum Information to Reference an External Ontology Term. In ICBO: International Conference on Biomedical Ontology pp. 87-90, 2009. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.371.6050&rep=rep1&type=pdf#page=101

[6] S.J.D. Cox, An explicit OWL representation of ISO/OGC Observations and Measurements. In O. Corcho, C. Henson, P. Barnaghi (Eds.), Proc. 6th Int. Work. Semant. Sens. Networks Co-Located with 12th Int. Semant. Web Conf. (ISWC 2013), Sun SITE Central Europe, Sydney, Australia, October 22nd, 2013, pp. 1–18, 2013. http://ceur-ws.org/Vol-1063/paper1.pdf

[7] S. Cox, Ontology for observations and sampling features, with alignments to existing models. Semantic Web Journal., in review. http://www.semantic-web-journal.net/content/ontology-observations-and-sampling-features-alignments-existing-models

[8] M. d'Aquin, A. Schlicht, H.,Stuckenschmidt, and M. Sabou, Ontology Modularization for Knowledge Selection: Experiments and Evaluations. In R. Wagner, N. Revell, & G. Pernul (Eds.), Database and Expert Systems Applications. Springer Berlin Heidelberg, pp. 874–883, 2007. http://dx.doi.org/10.1007/978-3-540-74469-6_85

[9] R.E., Duerr, et al. Formalizing the semantics of sea ice. Earth Science Informatics, 8(1): 51-62, 2015. http://doi.org/10.1007/s12145-014-0177-z

[10] European Commission, Open Data Portals. 2015. http://ec.europa.eu/digital-agenda/en/open-data-portals

[11] P.Fox and D.L. McGuinness, TWC Semantic Web Methodology, 2008.

http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology
[12] P. Fox, D.L. McGuinness, L. Cinquini, P. West, J. Garcia, , J.L. Benedict, and D. Middleton, Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience. Computers & Geosciences, 35(4): 724–738, 2009.. http://doi.org/10.1016/j.cageo.2007.12.019
[13] G. Guizzardi, Theoretical Foundations and Engineering Tools for Building Ontologies as Reference Conceptual Models. Semantic Web Journal, 1(1/2): 3-10, 2010. http://www.semantic-web-journal.net/content/theoretical-foundations-and-engineering-tools-building-ontologies-reference-conceptual-model
[14] H. Halpin, Social Semantics: The Search for Meaning on the Web. New York: Springer, 2013
[15] H. Halpin, P.J. Hayes, J.P. McCusker, D.L. McGuinness, and H.S. Thompson, When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In Volume 6496 of the series Lecture Notes in Computer Science: 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I, pp. 305-320, 2010. http://dx.doi.org/10.1007/978-3-642-17746-0_20
[16] R. Hoekstra, The Knowledge Reengineering Bottleneck. Semantic Web Journal, 1(1/2): 111-115, 2010. http://www.semantic-web-journal.net/content/knowledge-reengineering-bottleneck
[17] A. Hogan, A. Harth, , and A. Polleres, SAOR: Authoritative Reasoning for the Web. In J. Domingue & C. Anutariya (Eds.), The Semantic Web, Vol. 5367, pp. 76–90, 2008. Berlin, Heidelberg: Springer. http://doi.org/10.1007/978-3-540-89704-0_6
[18] E. Hyvönen, Preventing Interoperability Problems Instead of Solving Them. Semantic Web Journal, 2010. http://www.semantic-web-journal.net/content/preventing-interoperability-problems-instead-solving-them
[19] F.T. Imam, S. Larson, J.S. Grethe, A. Gupta, A. Bandrowski, and M.E. Martone, Development and Use of Ontologies inside the Neuroscience Information Framework: A Practical Approach. Frontiers in Genetics, 3: 00111, 2012. http://doi.org/10.3389/fgene.2012.00111
[20] K. Janowicz, S. Scheider, T. Pehle, and G.Hart, Geospatial Semantics and Linked Spatiotemporal Data – Past, Present, and Future. Semantic Web Journal, 3(4), 2012. http://www.semantic-web-journal.net/content/geospatial-semantics-and-linked-spatiotemporal-data-%E2%80%93-past-present-and-future
[21] H. Khan, B. Caruso, J. Corson-Rikert, D. Dietrich, B. Lowe, and G. Steinhart, DataStaR: Using the Semantic Web approach for Data Curation. The International Journal of Digital Curation, 6(2), 2011. http://ijdc.net/index.php/ijdc/article/view/192/257
[22] A. Krisnadhi, et al, An Ontology Pattern for Oceanographic Cruises: Towards an Oceanographer's Dream of Integrated Knowledge Discovery. In: Tom Narock and Peter Fox (eds.), The Semantic Web in Earth and Space Science: Current Status and Future Directions. Studies on the Semantic Web, IOS Press, Amsterdam, pp. 256-284, 2015.
[23] A.A. Krisnadhi, et al. The GeoLink Modular Oceanography Ontology. In: Proceedings ISWC2015, in press.
[24] X. Ma, Y. Chen, H. Wang, J. Erickson, P. West, and P. Fox, Deep Carbon Virtual Observatory: A cyber-enabled platform for linked science. SciDataCon 2014: International Conference on Data Sharing and Integration for Global Sustainability, 2014. http://www.researchgate.net/profile/Xiaogang_Ma3/publication/269990619_Deep_Carbon_Virtual_Observatory_A_cyber-enabled_platform_for_linked_science/links/54a6641e0cf257a63608f19c.pdf
[25] X. Ma, P. Fox, C. Tilmes, K. Jacobs, and A. Waple, Capturing provenance of global change information. Nature Climate Change, 4: 409-413, 2014. http://doi.org/10.1038/nclimate2141
[26] X. Ma, et al, Ontology engineering in provenance enablement for the National Climate Assessment. Environmental Modelling & Software, 61: 191–205, 2014. http://doi.org/10.1016/j.envsoft.2014.08.002
[27] D.L. McGuinness, Ontologies Come of Age. In D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster (Eds.), Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, Cambridge MA: MIT Press, pp 171-196, 2003.
[28] S. Mitchell, et al, Aligning research resource and researcher representation: the eagle-i and VIVO use case. In ICBO: International Conference on Biomedical Ontology, pp. 260-262, 2011. http://ceur-ws.org/Vol-833/paper42.pdf
[29] J.Z. Pan, L. Serafini,and Y. Zhao, Semantic Import: an Approach for Partial Ontology Reuse. In Proceedings of the 1st Workshop on Modular Ontologies (WoMO'06), 2006. http://ceur-ws.org/Vol-232/paper6.pdf
[30] F. Probst, et al. Connecting ISO and OGC Models to the Semantic Web. in Third International Conference on Geographic Information Science. Adelphi, MD, pp. 181-184, 2004.
[31] R.G. Raskin and M.J. Pan, Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). Computers & Geosciences, 31(9): 1119–1125, 2005. http://doi.org/10.1016/j.cageo.2004.12.004
[32] M.A. Rodriquez, J. Bollen, and H. Van de Sompel, (A Practical Ontology for the Large-Scale modeling of Scholarly Artifacts and their Usage. JCDL'07, June 18–23, 2007.
[33] G. Rust, Ontologyx. In Functional Requirements for Bibliographic Records Workshop Proceedings, Dublin, OH, May 2005.
[34] D. Tirry, A. Crabbé, and T. Steenberghen, Publishing metadata of geospatial indicators as Linked Open Data: a policy-oriented approach. In: Connecting a Digital Europe through Location and Place. Proc. of the AGILE'2014 International Conference on Geographic Information Science, pp. 1-6. 2014. http://m.agile-online.org/Conference_Paper/cds/agile_2014/agile2014_135.pdf
[35] C. Torniai, S. Essaid, B. Lowe, J. Corson-Rikert, and M. Haendel, Finding common ground: integrating the eagle-i and VIVO ontologies, In: Proceedings of the 4th International Conference on Biomedical Ontology, pp. 46-49, 2013. http://ceur-ws.org/Vol-1060/icbo2013_submission_20.pdf
[36] M. Uschold, M. Healy, K. Williamson, P. Clark, and S. Woods, Ontology reuse and application. In Proceedings of the International Conference on FOIS'98, 1998.
[37] A. Wilson, M. Cox, D. Elsborg, D. Lindholm, and T. Traver, A semantically enabled metadata repository for scientific data. Earth Science Informatics, 2014. http://dx.doi.org/10.1007/s12145-014-0175-1